SOFTWARE

DISEÑO DE UN AGREGADOR PARA LA GESTIÓN DE LOS *BIG DATA* INFORMATIVOS

Design of an aggregator for managing informative big data

Manuel Blázquez-Ochando

Manuel Blázquez-Ochando es profesor del *Departamento de Biblioteconomía y Documentación* de la *Universidad Complutense de Madrid* (*UCM*). Su investigación se centra en el desarrollo de software documental, agregadores de contenidos, algoritmos de recuperación, programas *web crawler* y técnicas de *scraping*. Es autor del buscador *WauSearch* y de la distribución de utilidades para bibliotecas, archivos y museos *AMPdoc*. Su actividad docente aborda la automatización de unidades de información, arquitectura de la información y metadatos, diseño de bases de datos y aplicaciones web documentales.
http://www.mblazquez.es
http://orcid.org/0000-0002-4108-7531

Universidad Complutense de Madrid, Facultad de Ciencias de la Documentación
Departamento de Biblioteconomía y Documentación
Santísima Trinidad, 37. 28010 Madrid, España
manublaz@ucm.es

Resumen

El artículo explica el diseño y características de un sistema de agregación de contenidos de código abierto. Entre las especificaciones del programa destaca un motor de procesamiento colaborativo, capacidades de monitorización de la recuperación de información en tiempo real, configuración del comportamiento del agregador, clasificación automática, alertas de noticias filtradas y nuevos tipos de representación de la información a partir de mapas relacionales interactivos. Por otra parte el programa de agregación está diseñado para gestionar miles de canales de sindicación en formato RSS. Además proporciona estadísticas que pueden servir para estudiar la producción informativa de cualquier sujeto productor y el impacto de la información publicada en otras fuentes. El resultado obtenido se traduce en módulos capaces de comparar las relaciones entre distintas informaciones o fuentes, determinar su grado de influencia, repercusión, impacto y cuantificación.

Palabras clave

Agregadores; *Big data*; Producción informativa; Canales de sindicación RSS; Recuperación de información.

Abstract

The design and characteristics of a new open source content aggregation program, AXYZ, are described. Several features of the program standout, including the processing engine of syndication channels, monitoring capability of information recovery in real time, possibility of configuration of the aggregator behavior, automatic content classification, and new models for representing information from relational interactive maps. On the other hand, the aggregation program is designed to manage thousands of syndication channels in the RSS format. It also provides statistics that can be used to study the production of any information producer and the impact of the information published in other sources. The AXYZ modules are capable of comparing the relationship between news or information from different sources and the degree of influence which is detected by patterns.

Keywords

Aggregators; Big data; Information production; RSS feeds; Information retrieval.

Blázquez-Ochando, Manuel (2016). "Diseño de un agregador para la gestión de los *big data* informativos". *El profesional de la información*, v. 25, n. 4, pp. 671-683.

http://dx.doi.org/10.3145/epi.2016.jul.17

Artículo recibido el 20-02-2016
Aceptación definitiva: 03-07-2016

1. Introducción

La creciente producción de información y su disponibilidad en la Red son el origen del concepto *big data*. **Mayer-Schönberger** (2013) lo define centrándose en dos aspectos: el almacenamiento de datos a gran escala y los métodos de análisis mediante patrones, con los que se advierten repeticiones y tendencias que sirven para crear modelos predictivos en campos tan diversos como la medicina, meteorología, criminología, finanzas, consumo o genómica, entre otros.

También se incide en el estudio de las correlaciones entre variables, para descubrir relaciones ocultas entre la información y datos masivos previamente almacenados, que pasarían inadvertidas. Con ello se consigue un mayor conocimiento de las causas y efectos que afectan a un objeto de estudio dado, todo lo cual revierte en ventajas competitivas. Tomando como punto de partida esta definición, se plantea la creación de un programa de agregación de contenidos, capaz de trabajar con textos informativos procedentes de miles de medios de comunicación, para su posterior filtrado y análisis correlacional.

Para abordar el objeto de estudio, se analiza la bibliografía científica sobre *big data* en relación con las aplicaciones de agregación en medios de información y comunicación y su tecnología. Una vez revisadas las últimas investigaciones, se propone un pliego de especificaciones con las características deseables para el diseño del nuevo programa de agregación, sus módulos, sets de datos y desarrollo. Finalmente se aborda el funcionamiento y sus características principales.

La búsqueda de artículos para la revisión del estado del arte, se ha ceñido a las bases de datos y buscadores: *Google Scholar*, *Web of Science* y *Scopus*.

La fórmula empleada en la búsqueda avanzada en Google Scholar es la siguiente:

> (intitle:"aggregator" AND intext:"RSS") OR (intitle:"RSS reader" AND intext:"aggregator") OR (intitle:"feed reader" AND intext:"aggregator") OR (intitle:"feed aggregator" AND intext:"RSS") OR (intitle:"RSS crawler" AND intext:"feed") OR (intitle:"syndication" AND intext:"feed").

2. *Big data* y agregación en periodismo y comunicación

Una cuestión recurrente cuando se aborda el tema de *big data* corresponde a los métodos de gestión, procesamiento y extracción de los datos. Según **Bansal** y **Kagemann** (2015) aún deben mejorarse los procedimientos de almacenamiento para cada tipo de datos. Esto se debe a su heterogeneidad temática y tipológica. Para dificultar aún más la tarea, el método de extracción y su procesamiento varía en función de su contexto de aplicación, género, formato e incluso fuente de procedencia. Por ejemplo la metodología para el tratamiento de textos planos, estructurados, imágenes y otros documentos multimedia, requiere patrones de clasificación específicos que permitan reconocer expresiones, vocabulario técnico y marcas audiovisuales basadas en polígonos, o bien espectrogramas de audio, entre otros. Otros autores también coinciden en la necesidad de dominar la extracción y el almacenamiento de datos, como paso previo a estudios analíticos (**Chen**; **Zhang**, 2014), que aún se encuentran en una etapa inicial. De hecho todavía resulta difícil combinar los *big data* procedentes de fuentes y temáticas dispares, si no existe una guía de relaciones hipotéticas, construida a priori por el investigador. Sin embargo el periodismo puede aprovecharse de infraestructuras ya conocidas, tales como los programas de agregación de contenidos (**Bazargani**; **Brinkley**; **Tabrizi**, 2013) para estudiar las relaciones entre las noticias y contenidos publicados en la Web por los medios de comunicación.

Este enfoque permite hablar del concepto *big data* informativo, por aludir al tratamiento masivo de los canales de comunicación de prensa, radio y televisión digitales. Sin embargo existen dificultades inherentes al producto informativo a la hora de diferenciar la originalidad, calidad, valor e interés de la información (**Katakis** *et al.,* 2009). Ello requiere aún del análisis documental y periodístico del profesional, que deberá dominar las técnicas de minería de datos para realizar periodismo de investigación, para ayudar a desvelar las correlaciones entre eventos y hechos a priori inconexos (**Colle**, 2013). Esto facilita las actividades de curación de contenidos y repercute en la sociedad a través de medios y plataformas digitales (**Guallar**; **Leiva-Aguilera**, 2013). Todo ello justifica la necesidad de construir sistemas de agregación de contenidos orientados a los *big data*, a la clasificación de contenidos, sistemas de alerta y medición de impacto.

> "Los *big data* generados por los medios puede ser tratado por agregadores. La dificultad estriba en diferenciar el valor de los contenidos y su impacto"

También se plantean otros retos derivados como la representación de noticias en tiempo real, para crear una fotografía instantánea del interés social y mediático. De hecho las últimas investigaciones en el área centran su atención en la revisión de sucesos y eventos, así como sus repercusiones en los flujos informativos internacionales (**Severo**; **Beauguitte**; **Pecout**, 2015). Previamente otros investigadores también atendieron al análisis de las noticias de propagación global con el propósito de identificar las fuentes de información más trascendentales e influyentes (**Gallé**; **Renders**; **Karstens**, 2013). Incluso se han llevado a cabo métodos para conocer la relevancia de los contenidos con mayor interés en la opinión pública (**Thelwall**; **Prabowo**; **Fairclough**, 2006). Esto fue posible usando canales de sindicación y tecnologías de agregación, que han servido como pilares básicos en la elaboración de dichos trabajos. Quizá el ejemplo que mejor encarna los *big data* informativos es el trabajo de **Travers** *et al.* (2014) en el que se revisaron 10 millones de noticias obtenidas durante 8 meses de agregación. Su análisis desveló que la *Ley de Pareto* se cumple al identificar que una minoría de canales de sindicación producía la mayor parte de los contenidos recopilados. Su trabajo desveló cuáles eran los medios digitales más productivos e influyentes y con qué publicaciones se alcanzó más éxito. También ayudó

a proporcionar datos métricos y de caracterización de las fuentes de información, la extensión media de los artículos, los fenómenos de duplicación y replicación, estrategias de posicionamiento y la capacidad de propagación informativa.

A pesar del valor que los canales de sindicación pueden tener en investigaciones de *big data*, uno de los obstáculos que los investigadores pueden encontrar es su localización y extracción. Según **Reichert** *et al.* (2011) esto es un problema de minería de datos que puede resolverse combinando estrategias de consulta en buscadores y programas de tipo *webcrawler*. Sin embargo la creación de una colección de canales de sindicación para su estudio y análisis, lejos de ser una recopilación masiva, también debe basarse en factores de calidad y pertinencia. Esto es, que además de obtener datos masivos las fuentes de información sean relevantes y suficientemente reconocidas en un dominio temático concreto, que facilite la interpretación de los contenidos que sean recuperados.

> "Los agregadores deben incorporar métodos que ayuden a clasificar mejor la información y aumentar su escalabilidad"

La realidad de las tecnologías de la información en torno a los *big data* y la agregación de contenidos han servido para crear nuevas empresas periodísticas, basadas en la reutilización de información que han cambiado el paradigma del sector (**Carlson**; **Usher**, 2015). Es posible competir con grandes grupos de comunicación, cuya adaptación al cambio es más lenta, situándolos en una posición de desventaja, con un adecuado empleo de los agregadores y la curación de contenidos. Un ejemplo de las consecuencias fue el caso Google News España (Barrio Garcés, 2014), un servicio que fue cerrado tras la aprobación de la reforma de *Ley 21/2014 de 4 de noviembre de propiedad intelectual* (*España*, 2014). La imposición del canon *AEDE* (*Asociación de Editores de Diarios Españoles*) a todos los portales de agregación de noticias con o sin ánimo de lucro, obligó a numerosas empresas a cesar sus actividades y a limitar las posibilidades de innovación en el sector. Todo ello fue debido en gran medida a la crisis de las empresas periodísticas que se vieron incapaces de competir ante los servicios de comparación de noticias y *newscliping*, que de forma gratuita se proporcionaban (**Guallar**, 2015). A pesar de todo la *Ley* no impide que los agregadores puedan ser usados o implementados en servidores locales (**López-Maza**, 2015). En torno al ámbito de las libertades cabe destacar que la transparencia se logra asegurando el libre acceso a la información y uno de los mejores medios que tiene la sociedad siguen siendo los agregadores de contenidos y la tecnología de sindicación (**Marty** *et al.*, 2010). Para alcanzar ese fin se tiene que asegurar un flujo constante de información, procedente de los medios, gobiernos y administraciones públicas. La sociedad tiene derecho a cotejar y contrastar la información de forma automática con miles de fuentes de información, para intentar conformar una opinión crítica, exhaustiva, fiel a la realidad y debidamente documentada (**Leaver**; **Willson**; **Balnaves**, 2012).

3. Tecnologías de agregación

El desarrollo de agregadores surge de la necesidad de controlar y procesar la información que generan millones de autores y editores en la Web. Si se asume que cada sitio web basado en CMS genera al menos un canal de sindicación, la cifra estimada superaría los 300 millones (*BuiltWith*, 2016). A esta cantidad habría que añadir los canales generados por blogs y redes sociales, obteniendo un volumen difícilmente abarcable. Una forma de afrontar el problema son los métodos de análisis multidimensional que tratan de procesar los *big data* en sets de datos no necesariamente caracterizados, para su posterior análisis (**Cuzzocrea**, 2015).

Dos de los programas más reconocidos en la materia son *Apache Hadoop* y *Apache Hive*. Su objeto es la agregación y análisis de cualquier tipo de contenido y su tratamiento a gran escala, de forma distribuida a través de una red de servidores. No podría considerarse un sistema específico para el tratamiento de canales de sindicación, aunque sí para el desarrollo de buscadores y programas *webcrawler*.

Otro enfoque para resolver el problema, corresponde a los métodos de agrupación de contenidos mediante algoritmos *K-mean* (**Li** *et al.*, 2007). Sin embargo su aplicación es compleja ya que los resultados que se obtienen pueden ser igualmente inabarcables. Incluso con un número relativamente bajo de canales de sindicación se pueden producir miles de grupos de noticias por efecto de una excesiva granularidad. Una solución a éste problema es el método de agrupación multimodal de las noticias (**Messina**; **Montagnuolo**, 2009). Esto implica usar el corpus textual para la agrupación de los contenidos, y con ello obtener un modelo estadístico-probabilístico que permita clasificar sucesivas noticias. Tal como se indica, resulta de gran utilidad para su aplicación en colecciones de textos con información heterogénea o rasgos semánticos bien caracterizados, lo que resulta muy interesante para este caso.

Otro método de clasificación es el referido al empleo de inteligencia artificial. Un agregador podría ser capaz de distinguir los artículos relevantes para el lector, basándose en su comportamiento (**Chen**; **Bøen**, 2008). Esto ayuda a eliminar los contenidos accesorios con un margen de error cercano al 20%. Otros investigadores también observaron el inconveniente de encontrar canales de sindicación no conocidos. Ello requería usar programas *web crawler* adaptados al caso, abriendo la investigación al concepto de agregadores híbridos, con funciones de lectura, clasificación y descubrimiento (**Lee** *et al.*, 2008).

Otro enfoque es concebir al agregador como un intermediario entre las fuentes de información y el suscriptor que recibe la información en un formato enriquecido, diferente a RSS (**O'Riordan**; **O'Mahoney**, 2011). Si bien el concepto representa ventajas a la hora de ampliar los metadatos originales, representa una redundancia en la multiplicación de los procesos de agregación. Esto se debe a que el suscriptor necesitaría otro programa de lectura diferente al intermediario. También se han creado proyectos de reutilización de código abierto, para poner en marcha agregadores en la plataforma *Wordpress* (**Isah**, 2012). Ello hace posible la creación de portales autoalimentados con noticias de terceros medios, con un bajo coste de mantenimiento.

Otro asunto abordado en la literatura científica es el rendimiento de los agregadores. Si se desea utilizar esta técnica en la investigación de los *big data*, se necesita mejorar sus capacidades y escalabilidad. Algunos autores afirman que la actualización periódica de los contenidos puede ser una solución para controlar la producción de información publicada (**Sia**; **Cho**; **Cho**, 2007). Sin embargo cuando las investigaciones requieren el análisis de varios miles de canales de sindicación, éste método puede provocar pérdidas de datos, al no poder abarcar a tiempo todo el volumen de noticias que se genera (**Horincar**; **Amann**; **Artières**, 2010).

En esta línea, se viene desarrollando el proyecto *RoSeS* (**Creus** *et al.,* 2011) que apuesta por un esquema de recuperación a gran escala, que radica en determinar tiempos de refresco para la revisión de los canales de sindicación, basándose en los factores del límite del ancho de banda, la cantidad de noticias publicadas y la frecuencia de actualización del canal. Ello determina un tiempo de revisión específico para cada canal de sindicación, ordenando su prioridad de procesamiento dentro de la lista que los agrupa. Esto se enmarca en un intento por predecir y caracterizar el comportamiento de la publicación de los contenidos sindicados (**Hmedeh** *et al.*, 2011) para crear mejores sistemas de agregación.

Otro caso es el proyecto *NectarRSS* que determina prioridades de procesamiento en los canales de sindicación, según las preferencias del usuario conforme a los contenidos consultados, su frecuencia y las palabras clave de sus búsquedas (**Samper** *et al.*, 2008). Ello favorece la concentración del esfuerzo de análisis en las fuentes que satisfacen sus necesidades de información por encima de las demás.

Todas las investigaciones parecen coincidir en la necesidad de mejorar los agregadores en torno a una serie de aspectos entre los que destacan:

- capacidad para filtrar la información, evitando el ruido y noticias irrelevantes para el lector: se recomienda el empleo de métodos de clasificación automática, autoaprendizaje y análisis estadístico del comportamiento del usuario en referencia a sus necesidades informativas;
- aumentar la escalabilidad para procesar más canales de sindicación: se proponen frecuencias de actualización periódica de las fuentes de información, reducir el régimen de actualización según estadísticas de producción, diseñar reglas predictivas según el comportamiento del usuario o bien establecer prioridades de procesamiento.

4. Metodología y desarrollo

Una vez revisada la bibliografía y el estado del arte en materia de *big data* informativos y agregación, se han planteado las siguientes fases metodológicas:

1) Definición de las especificaciones para diseñar el nuevo programa de *big data* informativos, teniendo en cuenta los aspectos sin resolver, advertidos en las investigaciones analizadas.

2) Diseño de un esquema modular que organice las funciones del programa conforme a cada especificación declarada.

3) Diseño de una base de datos que permita el almacenamiento eficiente de la información procedente de los medios de comunicación.

4) Elección del entorno de programación y construcción del agregador.

> “Una solución que mejora notablemente el rendimiento del agregador es la introducción de programas *parser* colaborativos. Esto ayuda a repartir la carga de trabajo y priorizar las fuentes más relevantes”

4.1. Especificaciones

Los puntos clave del programa se han basado en las conclusiones y experiencias publicadas por los investigadores en materia de *big data* informativos y diseño de programas de agregación. Se ha procurado el cumplimiento de las siguientes:

- diseñar un método que aumente la capacidad para procesar *big data* procedentes de canales RSS 2.0, clasificar o filtrar los contenidos relevantes y eliminar aquellos que superen un tiempo de archivo establecido;
- desarrollar un sistema de clasificación automática que permita equiparar los filtros definidos por el administrador y los contenidos recopilados en los canales de sindicación;
- introducir un buscador a texto completo en lenguaje natural con operadores de consulta exacta;
- elaborar métodos estadísticos que permitan conocer el número de canales de sindicación, fuentes de información, noticias activas y archivadas, cronología de publicaciones y contenidos clasificados;
- crear utilidades para el análisis de impacto y correlación de noticias con el fin de evaluar eventos y contenidos de interés social o mediático;
- hacer posible el uso del agregador fuera del dominio público de internet, para evitar la aplicación de la *Ley de propiedad intelectual*. Esto es lograr su funcionamiento autónomo con servidor y base de datos local en el equipo del usuario, sin que suponga dificultades de instalación y configuración;
- crear métodos alternativos de representación de la información que permitan observar la publicación de noticias en tiempo real o comprobar sus relaciones;
- concebir un método de monitorización activa de los procesos de análisis y recopilación de datos, para observar el desempeño del programa.

4.2. Diseño de módulos

En base a las especificaciones planteadas, el agregador *AXYZ* dispone de:

- área administrativa: formada por los módulos de configuración, mantenimiento, estadísticas, importación y exportación de datos, edición de canales de sindicación, filtros para la clasificación de contenidos y procesamiento o ejecución del programa;

Tabla 1. Módulos y funciones del agregador *AXYZ*

Módulo	Funciones	Módulo	Funciones
Configuración	Gestión de procesos	Filtros	Edición de filtros con operadores
	Opciones de archivo	Aplicación de análisis de impacto	Cálculo automático del impacto de una noticia
	Opciones de filtrado		Cálculo de correlación entre dos grupos de noticias
	Opciones de representación	Portada de noticias	Representación de noticias según relevancia de la fuente
	Opciones de alerta		Filtro de noticias por categorías y períodos temporales
Mantenimiento / Estado	Comprobación de errores	Tiempo real	Información recopilada en el momento en que se procesa
	Estado de la base de datos		Filtro de noticias por categorías y períodos temporales
	Reiniciar agregador	Noticias filtradas	Selección de últimas noticias filtradas
Estadísticas	Estadísticas generales		Impresión de dossiers según filtros
	Estadísticas personalizadas	Buscador	Consulta en lenguaje natural y frase exacta
Importación	Importar listas de canales de sindicación		Filtrar resultados según fuentes e intervalos temporales
	Importación OPML		
	Exportación OPML	Mapa de noticias	Mapa de noticias relacionadas según categorías
Edición	Edición masiva de canales de sindicación		
	Edición en detalle	Bloc de noticias	Guardar noticias seleccionadas por el lector
Procesamiento	Recopilación de noticias		Guardar noticias filtradas automáticamente
	Monitorización de procesos		Buscar entre las noticias guardadas
	5 hilos de ejecución *parser*		Filtrar noticias según fuentes e intervalos temporales

- área de representación de contenidos: presenta los módulos de estudio de impacto, mapa de noticias, noticias en tiempo real, portada de noticias según categorías temáticas, contenidos filtrados y bloc de noticias guardadas (tabla 1).

Este diseño permite controlar todas las fases de tratamiento de la información. El primer paso consiste en la configuración del programa para controlar la velocidad de recopilación de fuentes y su comportamiento ante errores, optimizar todas las noticias o solamente aquellas que fueron filtradas por el programa.

Una vez parametrizado, el módulo de importación permite cargar listas de enlaces correspondientes a canales de sindicación RSS, indicando su valor, clasificación temática e importancia. En el proceso se discriminan aquellos que no sean validados debido a errores de codificación, estructura o integridad, obteniendo al mismo tiempo los datos metadescriptivos de los que sí sean aceptados.

Una vez importados, el módulo de edición favorece la modificación puntual de los canales de sindicación para su preparación final.

Una vez dispuestos los canales, el módulo de filtros permite elaborar estrategias de filtrado con todos los descriptores que el documentalista requiera. Ello permite al programa clasificar las noticias antes de ser registradas en base de datos y facilita su posterior tratamiento estadístico.

Creados los filtros, el módulo de procesamiento pone en marcha el programa aplicando todos los parámetros previamente configurados.

Mientras *AXYZ* funciona, los módulos de representación de noticias, monitorización en tiempo real, mapa de noticias, buscador y bloc permanecen activos para proporcionar instrumentos que permitan realizar un seguimiento activo de los contenidos y publicaciones recuperadas.

4.3. Tablas y sets de datos

El programa *AXYZ* opera una base de datos *MySQL* con tablas y sets de datos diseñados a medida para registrar:

- canales de sindicación activos y temporales;
- noticias activas, archivadas, pendientes de eliminar y filtradas;
- categorías temáticas;
- filtros que se utilizarán para clasificar los contenidos:
- datos de control de ejecución.

A continuación se exponen las tablas que utiliza y una muestra del set de datos que ilustra el método de almacenamiento utilizado.

Tabla *categories*

Registra las categorías clasificatorias para los canales de sindicación de contenidos, de acuerdo con un máximo de tres bloques o facetas. Por ejemplo en el set de datos de la tabla 2 se identifica que la categoría "Prensa digital" está disponible en el bloque 1 y su número de orden es 100.

Tabla 2. Ejemplo de set de datos de la tabla *categories*

```
INSERT INTO 'categories' SET id='1', type='1', category='Prensa digital', value='100';
```

Tabla *control*

Registra el último canal de sindicación analizado, el número parcial de items analizados, el tiempo global de funcionamiento en formato *UNIX* y el identificador del próximo canal de sindicación. En el ejemplo de la tabla 3 se muestra que el

parser con identificador 1 acaba de analizar el canal de sindicación 4.309, recopilando un total de 39.440 noticias en 214 ciclos de análisis en un tiempo de 3 horas y 36 minutos. También se indica que el próximo canal de sindicación que se analizará tiene el identificador 1.075

Tabla 3. Ejemplo de set de datos de la tabla *control*

```
INSERT INTO 'control' SET id='1', lastFeed='4309', parItems='39440',
numLaps='214', timer='12960.5947565', start='1075';
```

Tabla *filters*

Contiene los datos y parámetros de los filtros creados por el usuario para clasificar las noticias automáticamente. Concretamente almacena el título del filtro, descripción, palabras y frases clave para ser usadas con operadores booleanos. La tabla 4 presenta el filtro 1 relativo a movilidad sostenible cuyas palabras clave se guardan con el formato original introducido por el usuario y en un formato depurado, separando los descriptores por plecas (barras verticales). Además se añade un campo destinado a indexación que recoge todos los términos clave del filtro para facilitar su recuperación.

Tabla 4. Ejemplo de set de datos de la tabla *filters*

```
INSERT INTO 'filters' SET id='34', title='Term. Transporte subterrÃ¡neo',
description='Noticias especializadas en metro', filter1and='Metro',
filter2and='metro', filter1or='suburbano, transporte, subterrÃ¡neo,
tÃºnel, estaciÃ³n', filter2or='suburbano|transporte|subterraneo|t
unel|estacion', filter1not='', filter2not='', filter1noise='Tren ligero',
filter2noise='tren ligero', indexer='transporte subterraneo metro
suburbano transporte subterraneo tunel estacion';
```

Tablas *feeds* y *feedstemp*

La tabla *feeds* registra los canales de sindicación activos, que van a ser usados. Contiene campos para la metadescripción del canal, la asignación de categorías, así como el número de identificación del programa *parser* que se encargará de procesar el canal. Otros datos de interés corresponden al valor del canal, fuente o dominio de procedencia y campo de indexación con los términos clave ya depurados y normalizados.

La tabla *feedstemp* es una copia de la tabla *feeds* y sirve para almacenar temporalmente los canales de sindicación durante el proceso de importación.

Tabla 5. Ejemplo de set de datos de la tabla *feeds*

```
INSERT INTO 'feeds' SET id='4449', regdated='2016-06-16 16:24:28',
title='Tendencias 21. Ciencia, tecnologÃa, sociedad y cultura',
link='http://www.tendencias21.net/xml/syndication.rss?r=190619',
description='Revista electrÃ³nica de ciencia, tecnologÃa, sociedad
y cultura. ISSN 2174-6850', language='es', copyright='', managingE-
ditor='', webMaster='', pubDate='Thu, 16 Jun 2016 16:00:01 +0200',
lastBuildDate='Thu, 16 Jun 2016 16:00:01 +0200', category='Science',
generator='', ttl='60', image='http://www.tendencias21.net/var/style/
logo.jpg', source='http:// tendencias21.net', cat1='1', cat2='3', cat3='5',
priority='1', core='5', value='90', indexer='tendencias ciencia tecnolo-
gia sociedad cultura revista electronica ciencia tecnologia sociedad
cultura issn 2174 6850 science';
```

Tablas *items* e *items365*

La tabla principal *items* almacena las noticias y contenidos de cada canal de sindicación. Recoge el identificador del filtro con el que se clasifica la noticia y el identificador del canal de sindicación en el que fue publicada. Otros datos fundamentales son el código *hash* de la noticia, su fecha de registro, título, enlace, descripción o contenido, autor, categoría temática, comentarios, archivos adjuntos o encapsulados, enlace permanente, fecha de publicación, dominio de la fuente, código fuente original, notas y campo de indexación.

La tabla *items365* es una copia de la tabla *items*, cuya función es almacenar las noticias que no han sido filtradas y que son susceptibles de eliminación automática, al cabo de un período de tiempo previamente configurado por el usuario. De esta forma el programa restringe el almacenamiento a los contenidos deseados, filtrados o seleccionados para su conservación.

Tabla 6. Ejemplo de set de datos de la tabla items

```
INSERT INTO 'items' SET id='61901', idfilter='34', idfeed='4449',
hash='52a838848f5140224013d0a427376cfe0efc0f4a', regda-
ted='2016-06-27 14:30:04', title='Los nuevos sistemas de metro ligero
son una soluciÃ³n sostenible y eficiente', link='http%3A%2F%2Fwww.
tendencias21.net%2FLos-nuevos-sistemas-de-metro-ligero-son-
una-solucion-sostenible-y-eficiente_a11267.html', description='Las
unidades de metro ligero de Ãºltima generaciÃ³n presentan inte-
resantes ventajas como soluciÃ³n para los sistemas de transporte
pÃºblico…', author='Pablo Javier Piacente', category='Transporte
sostenible', comments='', enclosure='', guid='www.tendencias21.
net%2C2016%3Arss-4110225', pubDate='Fri, 20 Apr 2012 09:45:00
+0200', source='http://tendencias21.net', htmlcontent='', notes='',
indexer='nuevos sistemas metro ligero solucion sostenible eficiente
unidades metro ligero ultima generacion presentan interesantes
ventajas solucion sistemas transporte publico…';
```


4.4. Programación del agregador

El entorno de desarrollo elegido para el agregador fue *Apache*, PHP y *MySQL* (AMP) por ser uno de los más extendidos y estandarizados. El método de programación empleado es de tipo estructurado y orientado a objetos. Esto significa que cada módulo dispone de un código de programación específico y unas rutinas que delimitan las opciones y funciones que se han predefinido en el apartado de especificaciones. Un ejemplo relevante de la metodología empleada, puede encontrarse en el algoritmo de procesamiento de noticias, encargado de revisar los canales de sindicación de los medios de comunicación, también conocido como *parser*. Su código estructurado permite inferir los siguientes pasos:

1) Cargar función de indexación, eliminación de etiquetas html y de iteración de tiempos para *benchmark* en tiempo real.

2) Selección de la base de datos.

3) Recuperación del identificador del último canal de sindicación analizado.

4) Liberación de memoria, antes de iniciar operaciones de recuperación.

5) Recuperación de los datos de la tabla control.

6) Calcular la fecha y hora de inicio y fin de ejecución.

7) Recuperación en tabla *feeds* del enlace al canal de sindicación que será consultado.

8) Envío de petición al servidor mediante funciones *cURL*.

9) Obtención del código fuente del canal de sindicación en formato XML.

10) Creación de un objeto DOM con el código fuente obtenido.

11) Selección de etiquetas contenedoras de las noticias mediante *XPath*.

12) Generación de *hash* identificativo de la noticia.

13) Comprobación de duplicados en base de datos usando el código *hash*.

14) Notación de tiempos de ejecución y número de noticias recopiladas para su supervisión.

15) Indexación del texto de las noticias.

16) Aplicación de filtros diseñados por el usuario.

17) Inserción de datos en la tabla *items*.

18) Eliminación de variables y liberación de memoria.

19) Iteración del proceso hasta la finalización de todos los canales de sindicación asignados al programa *parser*.

Un caso de programación orientada a objetos, corresponde al algoritmo de clasificación y filtrado. Su ejecución se produce únicamente cuando se declara su nombre clave y se transmiten las variables requeridas, devolviendo como resultado la clasificación positiva o negativa de las noticias. Su código a su vez es estructurado y se define con los siguientes pasos:

1) Recepción del texto depurado de la noticia.

2) Recuperación de los filtros definidos por el usuario y previamente almacenados en listas de tipo *array*.

3) Búsqueda de coincidencias parciales o totales de los términos del filtro conforme a la lógica booleana.

4) Cálculo de similaridad booleana en función de la frecuencia de aparición de los términos.

5) Almacenamiento del coeficiente de similaridad obtenido y cálculo del resto de filtros.

6) Selección del filtro con mejor coeficiente de similaridad, remitiendo su número de identificación.

7) En caso de que la noticia no sea clasificada, se devuelve el valor 0, que indica al programa que la noticia puede ser archivada.

> La posibilidad de calcular el impacto y la correlación entre grupos de noticias puede facilitar la investigación de tendencias periodísticas y sociales en los medios de comunicación

El resultado del trabajo se traduce en más de 20.000 líneas de código, distribuidas en 53 archivos originales. También se utilizaron 7 librerías externas para desempeñar las funciones de alerta por correo electrónico, generación de gráficos estadísticos, representaciones gráficas y rutinas de procesamiento de textos de tipo *wysiwyg* que se encuentran debidamente citadas en el código fuente del programa.

Terminada la fase de programación, el agregador fue instalado en la plataforma portable *Wamp Server2go* (**Haberkern, 2007**). Esto ha permitido que pueda ser utilizado

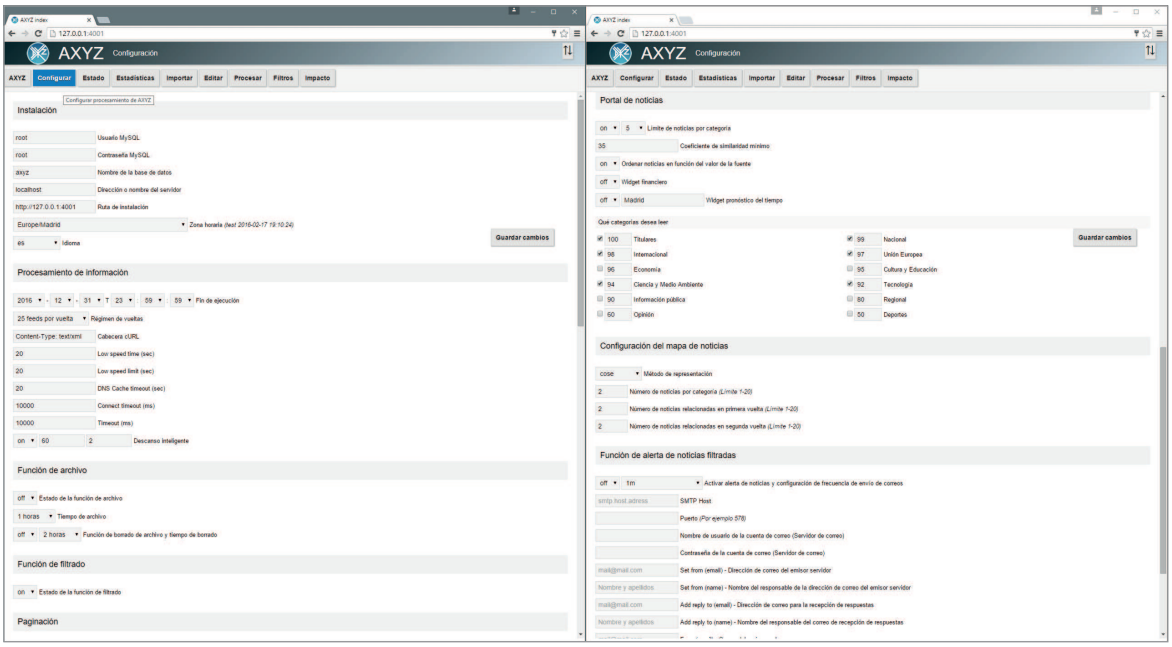

Figura 1. Módulo de configuración

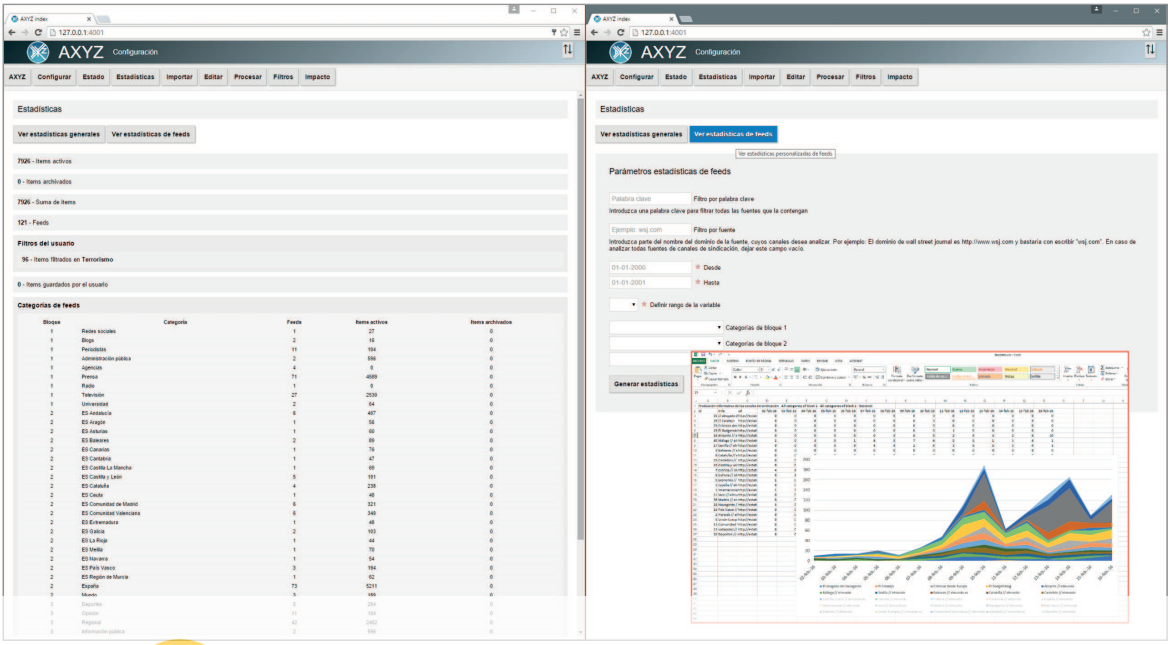

Figura 2. Opciones estadísticas del agregador

de forma local y autónoma sin necesidad de instalación o configuración. De hecho cualquier persona puede descargar el agregador en su ordenador y trabajar libremente como mejor considere, en el marco del derecho de acceso a la información pública, sin interferir en los dominios legales establecidos. El agregador, denominado *AXYZ*, se encuentra disponible para su experimentación, explotación científica y académica en el repositorio de software libre *SourceForge.net* en la dirección:

http://sourceforge.net/projects/axyznews

3. Características destacables del agregador *AXYZ*

El programa ha sido pensado para integrar todas las opciones y parámetros que controlan el funcionamiento de los módulos desde un módulo de configuración centralizado. Por ejemplo:

- temporizador de tareas
- progreso de la velocidad de recopilación de datos
- definición del método de archivo y expurgo
- filtrado de contenidos
- particularidades de la portada de noticias
- alertas informativas.

El módulo Estado supervisa las tablas de la base de datos, habilitando el mantenimiento, optimización o borrado de todos los contenidos recopilados, en caso de querer reiniciar la investigación en curso. En relación con las estadísticas, se proporcionan datos exactos sobre el número de contenidos activos, archivados, clasificados y distribuidos según categorías temáticas definidas por el usuario. También suministra datos sobre el número de canales de sindicación importados y su producción informativa según cronología, rangos temporales, fuente o palabras clave. La información puede ser exportada en hojas de cálculo para su posterior tabulación.

El módulo Importación se dedica a la inserción de listas de canales de sindicación, que pueden ser pre-clasificadas en torno a tres bloques de categorías temáticas, definidas por el administrador. También incorpora el concepto de valoración o importancia de las fuentes y la prioridad de procesamiento de las mismas. Por ejemplo un nivel de prioridad 1 forzará al agregador a revisar con más asiduidad los canales de sindicación. Si además el administrador valora la fuente, el programa ponderará positivamente las noticias que ésta produzca, ayudando a organizar mejor los resultados. También incorpora las opciones de importación y exportación en formato OPML (*outline processor markup language*), permitiendo archivar listas de canales de sindicación.

> El agregador *AXYZ* puede ser descargado libremente para usos académicos y científicos en: *http://sourceforge.net/projects/axyznews*

Cuando se importan canales de sindicación en el agregador, éstos son validados y puestos a su disposición en el módulo de edición. Resulta destacable la capacidad para editar bloques de 100 canales de sindicación, facilitando la modificación masiva de sus títulos, descripciones, fuentes, dominios, prioridad, importancia, fechas y temática. Otra ventaja se basa en la aparición de señales gráficas que avisan del estado de la clasificación de los registros, de forma que puedan ser localizados inmediatamente para su subsanación.

Resuelta la edición de los canales, el agregador puede ponerse en marcha usando la función de Procesamiento. Esta función abre una ventana de monitorización estratigráfica de las tareas de recopilación de noticias en los canales de sindicación previamente registrados. Sin embargo presenta innovaciones que han multiplicado la capacidad del agrega-

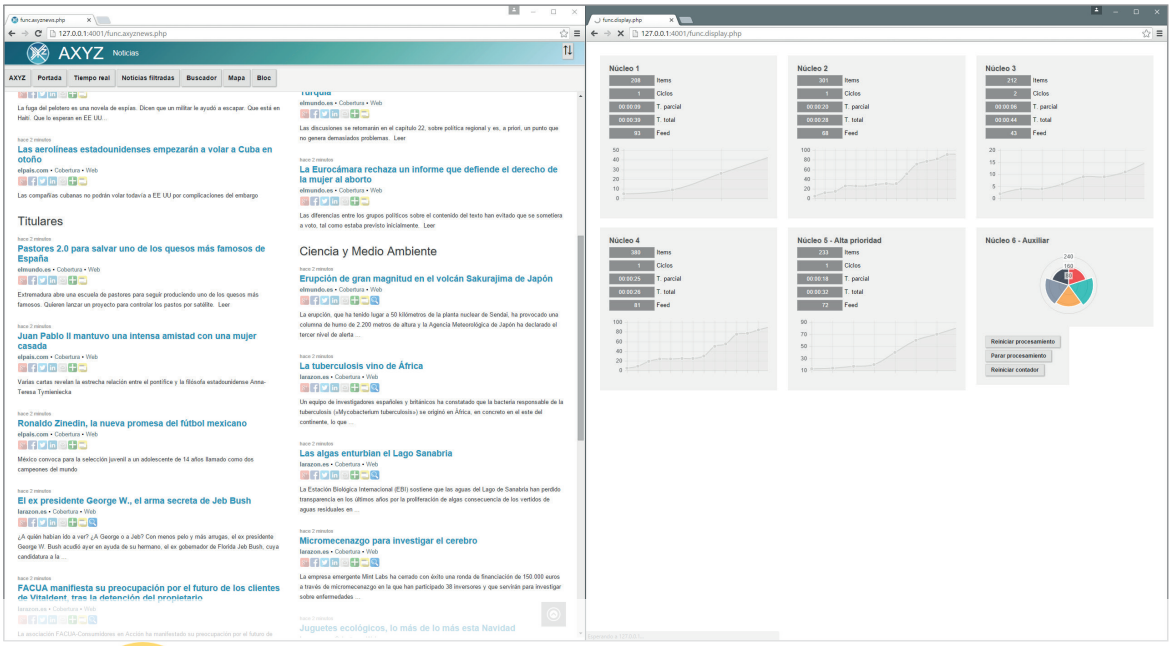

Figura 3. Portada de noticias y monitorización de la actividad del agregador

dor. Se trata de la introducción de 5 programas *parser* cuya ejecución independiente favorece la recuperación simultánea de las noticias actualizadas en los canales de sindicación. Esto se consigue asignando una lista de canales que será explorada por cada *parser* sin interferir en las operaciones de los demás. El trabajo colaborativo de los programas *parser* hace deseable la monitorización estadística de la actividad en tiempo real. Esto es, proporcionar al investigador información constante sobre:

- número de noticias recopiladas en cada *parser*;
- tiempo de ejecución parcial y total del proceso;
- número de ciclos o revisiones de los canales;
- número de identificación del canal que está siendo analizado.

> Se ha prestado especial atención a completar todas las funciones que cabría esperar en un agregador de contenidos, incluyendo buscadores, grafos de noticias, portada personalizable, edición de filtros y migración de datos

Todo ello favorece que el investigador pueda controlar mejor el programa, detectar errores, fallos y realizar pruebas de rendimiento, que permitan verificar el funcionamiento del agregador y su comportamiento.

El módulo Filtros ha sido diseñado para recoger todos los criterios de clasificación que serán utilizados para discriminar la información. El método de recuperación aplicado es el booleano, con los operadores de intersección, unión, negación y ruido. Cada filtro es registrado en base de datos y en un archivo independiente con el que se genera un patrón de consulta con expresiones regulares. Este detalle aumenta significativamente la eficiencia y capacidad de filtrado del agregador, que evita tener que emitir más consultas a la base de datos, obteniendo vía *NoSQL* los patrones que necesita para calcular la similaridad de las noticias en relación con los filtros.

Tabla 7. Ejemplo de *array NoSQL* para el filtrado de noticias

```
$arrayFilters[] = array(id => "34", filterAND => "metro", filterOR =>
"suburbano|transporte|subterraneo|tunel|estacion", filterNOT => "",
filterNOISE => "tren ligero");
```

Uno de los módulos de mayor utilidad para la investigación es Impacto. El programa comprende una interfaz para buscar las noticias recopiladas por el agregador y opciones para calcular el factor de impacto relativo y absoluto, según se explica en la tabla 8. Las fórmulas empleadas han sido diseñadas específicamente para poner en relación las noticias seleccionadas por el investigador con otras similares o bien con respecto a la totalidad. Esto permite conocer qué relevancia ha tenido un evento en los medios y qué cobertura absoluta ha representado en la producción informativa total. En cuanto al método para obtener noticias cuyo tema y contenido sea similar, se emplean consultas *SQL* a texto completo con el cuerpo de la noticia, usando su título y su clasificación o filtro si lo tuviera. El programa contabiliza en tal caso el número total de resultados cuyo coeficiente de similaridad sea superior a 20. Esta medida obliga a que los contenidos recuperados tengan un alto nivel de semejanza con respecto a la noticia seleccionada, asegurando un cálculo más preciso. Sin embargo, el investigador puede modificar estos valores para perfeccionar su funcionamiento o satisfacer necesidades de cálculo específicas.

Cabe indicar que el modelo matemático está inspirado en el cálculo del factor de impacto aplicable a las revistas científicas (**Garfield**, 2006). Sin embargo, a diferencia de éste, no incluye el factor temporal en el denominador de las fór-

Tabla 8. Factores de impacto y correlación calculados en el agregador *AXYZ*

$$Q = Qt_1 + Qt_2 + Qt_n = Nr \,\&\, Sr$$

La consulta *Q* con unos términos Qt_1, Qt_2 y Qt_n proporciona una serie de noticias relacionadas *Nr* y canales de sindicación relacionados *Sr*

$$N_{dr \to nu} = \sum Nr \to (Qtnu_1 \cup Qtnu_2 \cup Qtnu_n) \geq sim = 20$$

El número de noticias directamente relacionadas con la seleccionada por el usuario $N_{dr \to nu}$ es igual a la suma de noticias resultantes Σ *Nr*, de la consulta de los términos de dicha noticia ($Qtnu_1$ U $Qtnu_2$ U $Qtnu_n$), cuya similaridad sea igual o mayor que 20.

$$IFA_{nu} = \frac{N_{dr \to nu}}{\sum N}$$

El factor de impacto absoluto de la noticia seleccionada por el usario IFA_{nu} es igual al número de noticias directamente relacionadas $N_{dr \to nu}$ fraccionado por el número total de noticias recopiladas en el agregador Σ *N*.

$$IFR_{nu} = \frac{N_{dr \to nu}}{Nr}$$

El factor de impacto relativo de la noticia seleccionada por el usuario IFR_{nu} es igual al número de noticias directamente relacionadas $N_{dr \to nu}$ fraccionado por el número total de noticias recuperadas en la consulta de origen *Nr*.

$$IFA_S = \frac{S_{dr \to nu}}{\sum S}$$

El factor de impacto absoluto entre fuentes IFA_S es igual al número de canales de sindicación directamente relacionados con la consulta del usuario $S_{dr \to nu}$, fraccionado por el número total de canales de sindicación registrados en el agregador Σ *S*.

$$IFR_S = \frac{S_{dr \to nu}}{Sr}$$

El factor de impacto relativo entre fuentes IFR_S es igual al número de canales de sindicación directamente relacionados con la consulta del usuario $S_{dr \to nu}$, fraccionado por el número de canales de sindicación *Sr* relacionados con la consulta original *Q*.

mulas, ya que la frecuencia de publicación en los medios de comunicación es muy elevada. Por ello su aplicación se aconseja en rangos cronológicos breves de horas, días, semanas o pocos meses.

También se ha programado un método para calcular la correlación aproximada entre dos grupos de noticias filtradas. El algoritmo tiene en cuenta los siguientes pasos:

1) El investigador selecciona una noticia A, que será comparada y correlacionada con otra denominada B.

2) Extracción de las noticias similares en función de su coeficiente de similaridad mínimo.

3) Creación de listas de noticias similares para A y B.

4) Cálculo del coeficiente de similaridad de cada noticia de la lista A para cada noticia de la lista B.

5) Obtención de listas de coeficientes de similaridad entre el grupo de noticias de A con respecto a B y viceversa.

6) Aplicación de la fórmula de correlación de Pearson para calcular la correlación entre dos variables que corresponden a los coeficientes de similaridad obtenidos por el grupo de noticias de A y B. El resultado obtenido se puede exportar en formato CSV para su tratamiento estadístico y con ello obtener un diagrama de dispersión que permita determinar el grado de correlación obtenido y su desviación típica. Esta información ayudaría a los investigadores a trabajar con grandes cantidades de datos y demostrar que se suceden correlaciones entre diversos eventos y contenidos publicados.

Otros módulos de interés son los vinculados con las funciones de representación y consulta como la Portada de noticias y el Buscador. En ambos casos se proporcionan métodos de filtrado según categoría temática, intervalos temporales y dominio. En este marco también se encuentra el módulo de noticias en tiempo real, que muestra las últimas 15 informaciones recopiladas con un intervalo de refresco de 30 segundos.

También es destacable el módulo Mapa de noticias, ya que hace posible la creación de un grafo arborescente o circular, con las noticias relacionadas de cada categoría temática. En su desarrollo se ha utilizado la librería *Cytoscape.js* (**Franz** *et al.*, 2016) diseñada para la representación de redes. En este caso *Cytoscape* ha sido adaptado a la representación de las noticias, mostrando el título y el enlace de las mismas, así como su relación por afinidad. El criterio de selección usado por el programa es el grado de actualidad de las noticias, eligiendo siempre las más recientes y el coeficiente de similaridad basado en la materia, textos y titulares recuperados. Finalmente el agregador tiene la capacidad de generar dossiers de noticias filtradas y almacenar las noticias favoritas del lector, incluyendo también en este caso un buscador específico.

6. Conclusiones

El agregador *AXYZ* cumple los objetivos de la investigación aportando soluciones para calcular la repercusión e impacto mediático de la información, añadiendo funciones estadísticas personalizadas que cuantifican la producción informativa en las fuentes y canales de sindicación registrados. Por otra parte también es posible obtener el coeficiente de correlación entre dos eventos y sus noticias próximas.

En cuanto a rendimiento, el agregador proporciona un sistema de 5 programas *parser* colaborativos, que ayudan a redistribuir el trabajo de recopilación de contenidos y con ello aumentar la velocidad y capacidad de procesamiento. Además habilita la supervisión de su funcionamiento, a través de una ventana de monitorización que suministra datos y gráficas estadísticas en tiempo real. Ello contribuye a detectar errores si los hubiere y también a medir su rendimiento. Otras funciones adaptadas a los *big data* son los módulos de edición y de clasificación automática, diseñados para tratar bloques de registros y reducir la concurrencia de consultas en la base de datos.

La clasificación automática de contenidos se aplica al mismo tiempo que son recopilados en sus respectivas fuentes. El agregador aplica todos los filtros definidos por el administrador a

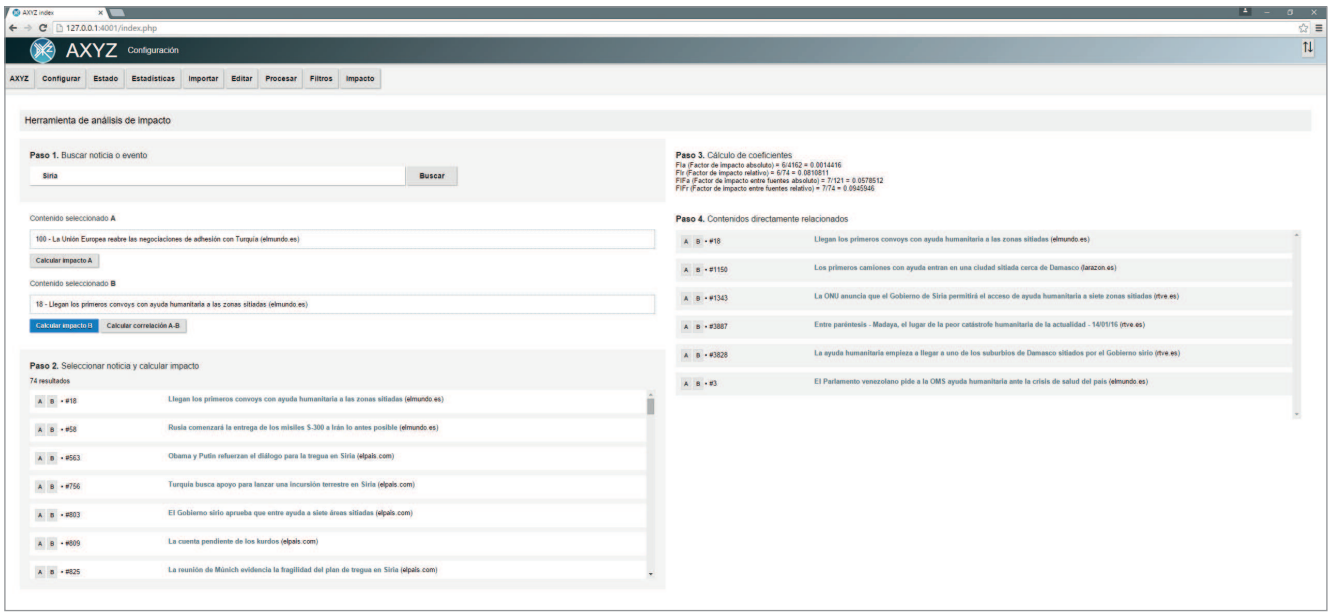

Figura 4. Módulo para el cálculo de impacto y correlación entre noticias

través de expresiones regulares, que combinan los términos y palabras clave con el método booleano de recuperación. A diferencia de otros programas, *AXYZ* genera un archivo *NoSQL* con las expresiones de consulta, al que se acude para clasificar la información. Esto evita tener que recurrir a la base de datos por cada noticia que se recuperada, lo que supone un ahorro del 50% en el número de peticiones al servidor.

El programa ha sido creado con todos los módulos que cabría esperar en un agregador de contenidos. Además incorpora otros menos comunes como:

- grafo de noticias relacionadas
- vista de noticias en tiempo real
- generador de cobertura informativa
- métodos de consulta predefinida en buscadores.

Las limitaciones que plantea la *Ley 21/2014 de 4 de noviembre de propiedad intelectual* (*España*, 2014) constituyen una amenaza para el desarrollo de tecnologías basadas en sindicación de contenidos. Afortunadamente el agregador *AXYZ* ha sido diseñado para funcionar en entorno local, lo que garantiza el derecho de acceso a la información publicada en medios de comunicación. Ello es debido a que la ley mantiene la excepción del derecho de copia privada permitiendo el uso de agregadores en el entorno particular, científico y académico. Cualquier investigador puede descargar el agrega-

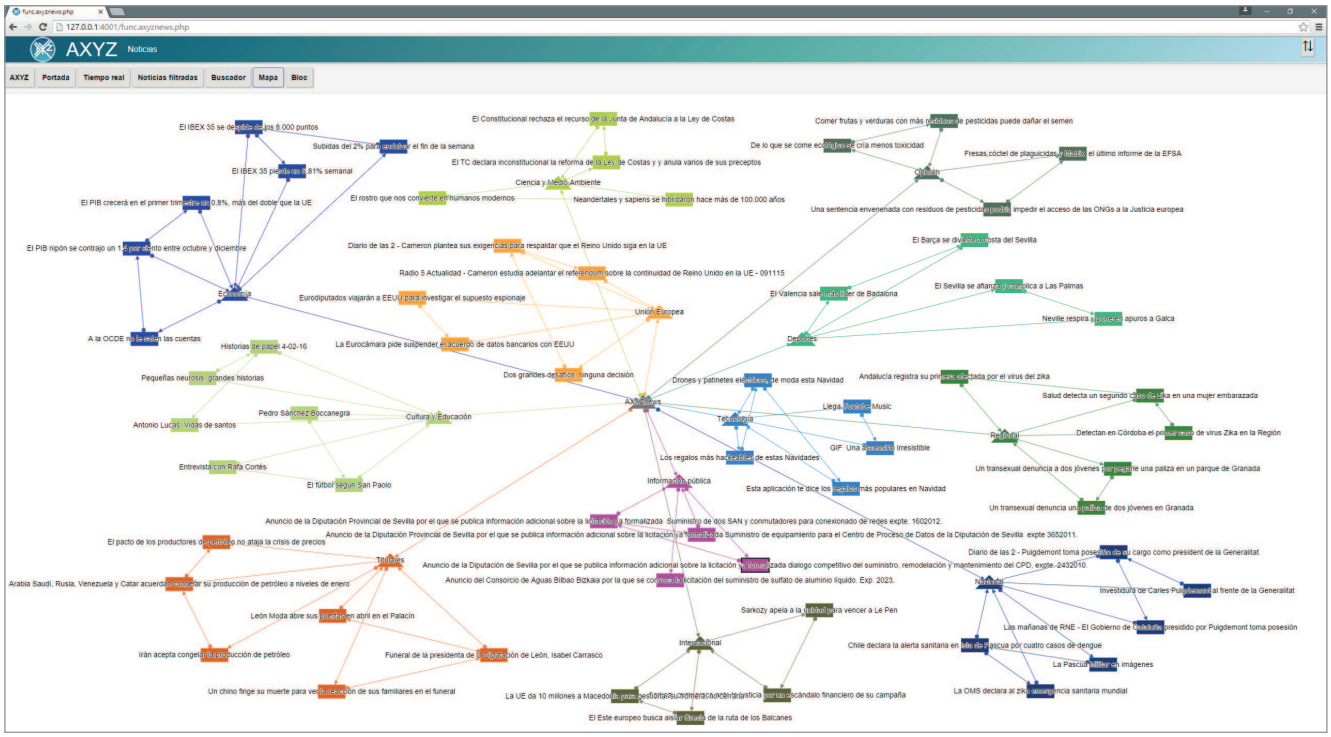

Figura 5. Mapa interactivo y relacional de noticias

dor y obtener un sistema de información equivalente a los extintos *Google Reader* y *Google News* (en España). Por otra parte puede conformar sus propias colecciones de fuentes, canales y noticias para iniciar proyectos de investigación informativa y documental de tipo métrico y cualitativo.

El agregador puede ser mejorado en sucesivas versiones atendiendo a la capacidad de inteligencia artificial para discernir las tendencias de consulta, corregir dinámicamente la relevancia de los contenidos basada en la experiencia del lector, ampliar las opciones de generación de grafos o enriquecer las noticias recopiladas con datos obtenidos a través de técnicas de rastreo web. También pueden mencionarse nuevas líneas de trabajo como el benchmarking comparativo entre agregadores, poner en práctica portales de canales de sindicación, análisis léxico, aplicaciones en otras áreas de conocimiento y adaptación del agregador para el intercambio de información bibliográfica o documental.

Bibliografía

Bansal, Srividya K.; **Kagemann, Sebastian** (2015). "Integrating big data: A semantic extract-transform-load framework". *Computer*, v. 48, n. 3, pp. 42-50.
http://doi.ieeecomputersociety.org/10.1109/MC.2015.76

Bazargani, Sahar; **Brinkley, Julian**; **Tabrizi, Nassehzadeh** (2013). "Implementing conceptual search capability in a cloud-based feed aggregator". En*: 3rd Intl conf on innovative computing technology (Intech)*, 29-31 Aug., pp. 138-143.
http://dx.doi.org/10.1109/INTECH.2013.6653631

BuiltWith (2016). *CMS usage statistics*. *Statistics for websites using CMS technologies*.
http://trends.builtwith.com/cms

Carlson, Matt; **Usher, Nikki** (2015). "News startups as agents of innovation: For-profit digital news startup manifestos as metajournalistic discourse". *Digital journalism*, v. 4, n. 5, pp. 1-19.
http://dx.doi.org/10.1080/21670811.2015.1076344

Chen, Philip; **Zhang, Chun-Yang** (2014). "Data-intensive applications, challenges, techniques and technologies: A survey on big data". *Information sciences*, v. 275, pp. 314-347.
https://goo.gl/kQJMSH
http://dx.doi.org/10.1016/j.ins.2014.01.015

Chen, Weiqin; Bøen, Torbjørn (2008). "A personalized RSS news filtering agent". En: Ellis, Richard; Allen, Tony; Petridis, Miltos. *Applications and innovations in intelligent systems XV*. Londres: Springer, pp. 321-326. ISBN: 978 1 84800 086 5
http://dx.doi.org/10.1007/978-1-84800-086-5_25

Colle, Raymond (2013). "Prensa y big data: el desafío de la acumulación y análisis de datos". *Mediterranean journal of communication*, v. 4, n. 1, pp. 275-282.
http://dx.doi.org/10.14198/MEDCOM2013.4.1.13

Creus, Jordi; **Amann, Bernd**; **Travers, Nicolas**; **Vodislav, Dan** (2011). "RoSeS: A continuous content-based query engine for RSS feeds". En: *Procs of the 20th ACM Intl conf on information and knowledge management*, 2011, pp. 2549-2552.
http://cedric.cnam.fr/fichiers/art_2086.pdf
http://dx.doi.org/10.1145/2063576.2064016

Cuzzocrea, Alfredo (2015). "Aggregation and multidimensional analysis of big data for large-scale scientific applications: models, issues, analytics, and beyond". En*: Procs of the 27th Intl conf on scientific and statistical database management*, pp. 23.
http://dx.doi.org/10.1145/2791347.2791377

España (2014). "Ley 21/2014, de 4 de noviembre, por la que se modifica el texto refundido de la Ley de propiedad intelectual, aprobado por Real decreto legislativo 1/1996, de 12 de abril, y la Ley 1/2000, de 7 de enero, de enjuiciamiento civil". *BOE*, n. 268, 5 de noviembre, pp. 90404-90439.
https://www.boe.es/diario_boe/txt.php?id=BOE-A-2014-11404

Franz, Max; **Lopes, Christian T.**; **Huck, Gerardo**; **Dong, Yue**; **Sumer, Onur**; **Bader, Gary D.** (2016). "Cytoscape.js: a graph theory library for visualisation and analysis". *Bioinformatics*, v. 32, n. 2, pp. 309-311.
http://dx.doi.org/10.1093/bioinformatics/btv557

Gallé, Matthias; **Renders, Jean-Michel**; **Karstens, Eric** (2013). "Who broke the news?: an analysis on first reports of news events". En: *Procs of the 22nd Intl conf on World Wide Web companion*, pp. 855-862.
http://www2013.org/companion/p855.pdf
http://dx.doi.org/10.1145/2487788.2488066

Garfield, Eugene (2006). "The history and meaning of the journal impact factor". *Jama*, v. 295, n. 1, pp. 90-93.
http://garfield.library.upenn.edu/papers/jamajif2006.pdf
http://dx.doi.org/10.1001/jama.295.1.90

Guallar, Javier (2015). "Prensa digital en 2013-2014". *Anuario ThinkEPI*, v. 9, pp. 153-160.
http://dx.doi.org/10.3145/thinkepi.2015.37

Guallar, Javier; **Leiva-Aguilera, Javier** (2013). *El content curator. Guía básica para el nuevo profesional de internet*. Barcelona: UOC. Colección El profesional de la información, n. 24. ISBN: 978 84 9064 018 0

Haberkern, Timo (2007). *Server2Go*.
http://www.server2go-download.de/download/server2go_a22_psm.zip

Hmedeh, Zeinab; **Vouzoukidou, Nelly**; **Travers, Nicolas**; **Christophides, Vassilis**; **Du-Mouza, Cedric**; **Scholl, Michel** (2011). "Characterizing web syndication behavior and content". En: Bouguettaya, Athman; Hauswirth, Manfred; Liu, Ling. *Web information system engineering (2011)*. Sydney: Springer, pp. 29-42. ISBN: 978 3 642 24434 6
http://cedric.cnam.fr/fichiers/art_2162.pdf
http://dx.doi.org/10.1007/978-3-642-24434-6_3

Horincar, Roxana; **Amann, Bernd**; **Artières, Thierry** (2010). "Best-effort refresh strategies for content-based RSS feed aggregation". En: Chen, Lei; Triantafillou, Peter; Suel, Torsten. *Web information systems engineering (2010)*. Hong Kong: Springer, pp. 262-270. ISBN: 978 3 642 17616 6
http://dx.doi.org/10.1007/978-3-642-17616-6_24

Isah, Haruna (2012). "Full data controlled web-based feed aggregator". *International journal of computer science & information technology*, v. 4, n. 3, pp. 71-84.
http://dx.doi.org/10.5121/ijcsit.2012.4307

Katakis, Ioannis; **Tsoumakas, Grigorios**; **Banos, Evangelos**; **Bassiliades, Nick**; **Vlahavas, Ioannis** (2009). "An adaptive personalized news dissemination system". *Journal of intelligent information systems*, v. 32, n. 2, pp. 191-212.
http://dx.doi.org/10.1007/s10844-008-0053-8

Leaver, Trama; **Willson, Michele**; **Balnaves, Mark** (2012). "Transparency and the ubiquity of information filtration?". *Ctrl-Z: New media philosophy*, v. 1, n. 2.
http://www.ctrl-z.net.au/articles/leaver-willson-balnaves-transparency-and-the-ubiquity-of-information-filtration

Lee, Bum-Suk; **Im, Jin-Woo**; **Hwang, Byung-Yeon**; **Zhang, Du** (2008). "Design of an RSS crawler with adaptive revisit manager". En: *SEKE*, 2008, pp. 219-222.
http://dblp.uni-trier.de/db/conf/seke/seke2008.html

Li, Xin; **Yan, Jun**; **Deng, Zhihong**; **Ji, Lei**; **Fan, Weiguo**; **Zhang, Benyu**; **Chen, Zheng** (2007). "A novel clustering-based RSS aggregator". En: *Procs of the 16th Intl conf on World Wide Web*, 2007, pp. 1309-1310. ISBN: 978 1 59593 654 7
http://www2007.org/posters/poster931.pdf
http://dx.doi.org/10.1145/1242572.1242824

López-Maza, Sebastián (2015). "El límite sobre agregadores y buscadores". En: Rodríguez-Cano, Rodrigo B. *La reforma de la Ley de propiedad intelectual*. Valencia: Tirant lo Blanch, pp. 89-111. ISBN: 978 84 9086 664 1

Marty, Emmanuel; **Rebillard, Franck**; **Smyrnaios, Nikos**; **Touboul, Annelise** (2010). "Variété et distribution des sujets d'actualité sur Internet. Une analyse quantitative de l'information en ligne". *Mots. Les langages du politique*, n. 93, pp. 107-126.
http://dx.doi.org/10.4000/mots.19832

Mayer-Schönberger, Viktor; **Cukier, Kenneth** (2013). *Big data: la revolución de los datos masivos*. Madrid: Turner. ISBN: 978 84 15427 81 0

Messina, Alberto; **Montagnuolo, Maurizio** (2009). "A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval". En: *Procs of the 18th Intl conference on World Wide Web*, 2009, pp. 321-330. ISBN: 978 1 60558 487 4
http://ra.ethz.ch/CDstore/www2009/proc/docs/p321.pdf
http://dx.doi.org/10.1145/1526709.1526753

O'Riordan, Adrian P.; **O'Mahoney, Oliver** (2011). "Engineering an open web syndication interchange with discovery and recommender capabilities". *Journal of digital information*, v. 12, n. 1.
https://cora.ucc.ie/handle/10468/980

Reichert, Sandro; **Urbansky, David**; **Muthmann, Klemens**; **Katz, Philipp**; **Wauer, Matthias**; **Schill, Alexander** (2011). "Feeding the world: a comprehensive dataset and analysis of a real world snapshot of web feeds". En: *Procs of the 13th Intl conf on information integration and web-based applications and services*, 2011, pp. 44-51. ISBN: 978 1 4503 0784 0
http://dx.doi.org/10.1145/2095536.2095546

Rodríguez-Cano, Rodrigo B. (2015). "Tasa *Google* o canon *AEDE*: una reforma desacertada". *Aranzadi civil-mercantil. Revista doctrinal*, v. 1, n. 11, pp. 53-94.

Samper, Juan J.; **Castillo, Pedro A.**; **Araujo, Lourdes**; **Merelo, Juan J.**; **Cordon, Oscar**; **Tricas, Fernando** (2008). "NectaRSS, an intelligent RSS feed reader". *Journal of network and computer applications*, v. 31, n. 4, pp. 793-806.
http://dx.doi.org/10.1016/j.jnca.2007.09.001

Severo, Marta; **Beauguitte, Laurent**; **Pecout, Hugues** (2015). "Archiving news on the Web through RSS flows. A new tool for studying international events". En: *Resaw. Web archives as scholarly sources: Issues, practices and perspectives*.
https://halshs.archives-ouvertes.fr/halshs-01187828

Sia, Ka-Cheung; **Cho, Junghoo**; **Cho, Hyun-Kyu** (2007). "Efficient monitoring algorithm for fast news alerts". *Knowledge and data engineering*, v. 19, n. 7, pp. 950-961.
http://dx.doi.org/10.1109/TKDE.2007.1041

Thelwall, Mike; **Prabowo, Rudy**; **Fairclough, Ruth** (2006). "Are raw RSS feeds suitable for broad issue scanning? A science concern case study". *Journal of the American Society for Information Science and Technology*, v. 57, n. 12, pp. 1644-1654.
http://dx.doi.org/10.1002/asi.20334

Travers, Nicolas; **Hmedeh, Zeinab**; **Vouzoukidou, Nelly**; **Du-Mouza, Cedric**; **Christophides, Vassilis**; **Scholl, Michel** (2014). "RSS feeds behavior analysis, structure and vocabulary". *International journal of web information systems*, v. 10, n. 3, pp. 291-320.
http://dx.doi.org/10.1108/IJWIS-06-2014-0023